Spring 2025 Optimal Transport Final Project: Wasserstein Gradient Flow and Otto Calculus

Minji Kim, Hyeon Lee

December 12, 2025

1 Introduction

Gradient flows arise in the context of minimization (or maximization) of a function $F : X \to \mathbb{R}$ on some space X. The name "gradient flow" suggests that the object of interest is a curve $(x_t)_{t\in[0,1]}$ in X parametrized by time t, which *flows* along the direction of (the negative of) the gradient of F i.e. the steepest descent direction. Gradient flows are widely understood in the case of $X = \mathbb{R}^d$ for some d. Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is an objective function and x_0 be an initial point. Then the gradient flow, $(X_t)_{t\in[0,1]}$, is described by the following partial differential equation:

$$\partial_t X_t = -\nabla f(X_t) \ {}^\forall t \geq 0, \quad X_0 = x_0. \tag{1}$$

The curve (X_t) flows towards a valley of the hypersurface $\{(x, f(x)) : x \in X\}$ and converges at least to a local minimum. The popular gradient descent algorithms can be considered as discretized gradient flows. The theory of gradient flows can be developed in a broader context where the underlying space X is a metric space, which is not necessarily a vector space, when some additional tools for calculation are available. Otto calculus is such a tool for gradient flows in Wasserstein spaces, namely Wasserstein gradient flows.

An interesting and illustrative example to which Wasserstein gradient flow is applied is Gaussian variational inference (VI). Given a target distribution π on $\mathbb{R}^d$, Gaussian VI aims to estimate a Gaussian distribution q that is closest to π under Kullback-Leibler divergence. Let us denote by $\mathrm{BW}(\mathbb{R}^d)$ the Bures-Wasserstein space, which is the collection of all non-degenerate Gaussian distributions on $\mathbb{R}^d$ (equipped with 2-Wasserstein distance). Gaussian VI is formulated as follows.

$$q^\star = \underset{q\in\mathrm{BW}(\mathbb{R}^d)}{\arg\min}\ KL(q\|\pi). \tag{2}$$

A popular approach to this problem is using Evidence Lower Bound (ELBO) as a surrogate for KL divergence, which has a simpler form and is easier to optimize. Wasserstein gradient flow offers an alternative, direct approach to this problem. It turns out that KL divergence is geodesically convex in its first argument for proper choice of the target distribution π, hence exponential convergence rate is guaranteed.

This report consists of three chapters. First, we review a general theory of Wasserstein gradient flows. A key ingredient to this is the continuity equation; it defines tangent spaces of Wasserstein spaces on which gradient vectors for Wasserstein spaces are defined. Next, we study geodesic convexity of KL divergence and arrive at exponential convergence rate of the gradient flow. The report concludes with an application to Gaussian variational inference with explicit expressions for the flow (μ_t).

For clarity of arguments, in this report, we restrict our discussion to $W_{2,ac}(\mathbb{R}^d)$, the space of probability measures on $\mathbb{R}^d$ that are absolutely continuous w.r.t. the Lebesgue measure equipped with the Wasserstein 2-distance. We remark that the theory of Wasserstein gradient flows exists for larger classes $\mathbb{W}_p(\mathbb{R}^d), 1 \leq p < \infty$.

2 The Continuity Equation

The first step for doing a calculus on Wasserstein spaces is to define a tangent vector of a curve γ in $W_{2,ac}(\mathbb{R}^d)$ at each $t \in (0,1)$. Literatures have revealed that the *continuity equation*, which we describe in this section, gives a useful notion of tangent vectors of curves in $W_{2,ac}(\mathbb{R}^d)$.

The continuity equations comes from considering a distribution of tiny particles each of which follows a given time-varying velocity vector field

$$(v_t : \mathbb{R}^d \to \mathbb{R}^d)_{t\in[0,1]}.$$

A particle that was located at x_0 at time $t = 0$ moves according to the following partial differential equation.

$$\dot{x}_t := \frac{d}{dt} x_t = v_t(x_t). \tag{3}$$

More generally, if we denote by $X_t(x)$ the location at time t of a particle which was located at x at time $t = 0$, its dynamics is written as

$$\dot{X}_t = v_t(X_t). \tag{4}$$

Now given that the distribution of the particles at $t = 0$ was μ, we are interested in the distribution μ_t of particles at each time $t \in [0,1]$. The curve (μ_t) is described as the following set of equations.

$$\mu_t = (X_t)_{\#}\mu, \quad \dot{X}_t = v_t(X_t), \quad X_0(x) = x \quad \text{for} \quad t \in [0,1]. \tag{5}$$

A curve μ_t that evolves as in (5) can be directly described using v_t and not introducing the solution X_t. Consider a compactly supported, infinitely differentiable function $\varphi \in C_c^\infty(\mathbb{R}^d;\mathbb{R})$. This function is nice enough so that one can handily exchange differentiations and integrals. In particular,

$$\begin{aligned}
\int \varphi(x)\partial_t\mu_t(x)dx &= \partial_t\left(\int \varphi(x)\mu_t(x)dx\right) \\
&= \partial_t \mathbb{E}_{Z\sim\mu_t}\varphi(Z) \\
&= \partial_t \mathbb{E}_{Z\sim\mu}\varphi(X_t(Z)) \\
&= \partial_t \int \varphi(X_t(x))\mu(x)dx \\
&= \int \langle \nabla\varphi(X_t(x)), \dot{X}_t(x)\rangle \mu(x)dx \\
&= \int \langle \nabla\varphi(X_t(x)), v_t(X_t(x))\rangle \mu(x)dx \\
&= \int \langle \nabla\varphi(x), v_t(x)\rangle \mu_t(x)dx \\
&= -\int \varphi(x)\operatorname{div}(\mu_t v_t)(x)\mu_t(x)dx,
\end{aligned}$$

where the last equality follows the divergence theorem. Since the equality holds for any choice of φ, we get

$$\partial_t \mu_t + \operatorname{div}(\mu_t v_t) = 0. \tag{6}$$

We say a pair (μ_t, v_t) *solves the continuity equation* if they satisfy (6) on $\mathbb{R}^d$.

We have seen that a family of vector fields v_t generates a curve μ_t in $W_{2,ac}(\mathbb{R}^d)$ from a initial measure μ. Next question is if we can recover v_t from μ_t. The answer is for an absolutely continuous curve μ_t in $W_{2,ac}(\mathbb{R}^d)$ (in the general notion of absolute continuity of curves in metric spaces), there exists a family of vector fields v_t that generates mu_t, but such a family is not necessarily unique. Among those families, we choose the one with the minimum magnitude. Let us measure the magnitude of a vector field using μ_t as $\|v_t\|_{L^2(\mu)}$.

Theorem 2.1. *Let μ_t be an absolutely continuous curve in $W_{2,ac}(\mathbb{R}^d)$. There exists a unique vector field v_t such that*

(i) $(\mu_t, v_t)_{t\in[0,1]}$ solves the continuity equation, and

(ii) for any other family w_t of vector fields such that (μ_t, w_t) solves the continuity equation, $\|v_t\|_{L^2(\mu_t)} \leq \|w_t\|_{L^2(\mu_t)}$ for all $t \in [0, 1]$.

We define v_t the tangent vector of μ_t at t.

Why we define the vector field as a tangent vector, and why we measure its magnitude in the $L^2(\mu_t)$ space? The following theorem states that the Wasserstein distance is recovered from the integral of velocity of the curves connecting two points. In this way, the tangent vectors is really compatible with the Wasserstein metric.

Theorem 2.2 (Benamou-Brenier)**.** *For $\mu, \nu \in \mathbb{W}_2(\mathbb{R}^d)$,*

$$W_2^2(\mu,\nu) = \inf\left\{\int_0^1 \|v_t\|^2_{L^2(\mu_t)} dt : \partial_t\mu_t + div(\mu_t v_t) = 0, \mu_0 = \mu, \mu_1 = \nu\right\}.$$

3 Tangent Space of $W_{2,ac}(\mathbb{R}^d)$

Next we study the tangent space $T_\mu W_{2,ac}(\mathbb{R}^d)$ in more detail. The previous theorem suggests that the tangent vectors at μ are elements of $L^2(\mu)$. We will define the tangent space at μ as a subset of $L^2(\mu)$, which contains all tangent vectors that arise from the continuity equation, which is also a closed linear subspace of $L^2(\mu)$. We can explicitly find tangent vectors at μ from any curves μ_t from the following theorem.

Theorem 3.1 (Ambrosio et al., 2008, Theorem 8.4.6)**.** *Let μ_t be an absolutely continuous curve in $W_{2,ac}(\mathbb{R}^d)$ and (μ_t, v_t) solves the continuity equation. Then,*

$$v_t = \lim_{h\to 0+} \frac{T_{\mu_t \to \mu_{t+h}} - id}{h} \quad in \;\; L^2(\mu_t). \tag{7}$$

To motivate the definition of the tangent space, we restrict our attention to the curves of the a special form. Let ν be another measure and $T_{\mu\to\nu}$ be the optimal transport map from μ to ν. Define a curve

$$\mu_t = ((1-t)x + tT(x))_\# \mu, \quad t \in [0, 1].$$

Then by Brenier's theorem, for any $h \in (0,1)$, the optimal transport map from μ_0 to μ_t is given by

$$T_{\mu_0 \to \mu_h}(x) = (1-h)x + hT(x).$$

Therefore,

$$v_0(x) = \lim_{h \to 0+} \frac{T_{\mu_0 \to \mu_h}(x) - x}{h} = \lim_{h \to 0+} \frac{(1-h)x + hT(x) - x}{h} = T(x) - x.$$

Again, by Brenier's map, $T(x) = \nabla \varphi(x)$ for some convex function φ. It can be shown that tangent vectors from non-geodesic curves are approximated by tangent vectors of geodesic curves. This leads to the following definition.

Definition 3.2. *The tangent space of $W_{2,ac}(\mathbb{R}^d)$ at μ is defined as the following space.*

$$T_\mu W_{2,ac}(\mathbb{R}^d) = \overline{\{\lambda(\nabla\varphi(x) - x) : \varphi \text{ is convex}, \lambda \in \mathbb{R}\}}^{L^2(\mu)}.$$

In particular, the tangent space is a closed linear subspace of $L^2(\mu)$, and hence is a Hilbert space.

Below is a more handy representation of the tangent space.

Theorem 3.3 (Ambrosio et al., 2008, Theorem 8.5.1)**.** *The following two sets are the same.*

$$\overline{\{\lambda(\nabla\varphi - \mathrm{id}) : \varphi : \mathbb{R}^d \to \mathbb{R}^d \text{ is convex}, \lambda \geq 0\}}^{L^2(\mu)} = \overline{\{\nabla\varphi : \varphi \in C_c^\infty(\mathbb{R}^d)\}}^{L^2(\mu)}.$$

4 Wasserstein Gradient

Wasserstein gradient is defined by borrowing an idea from the gradient of a functional $f : \mathbb{R}^d \to \mathbb{R}$. For the Euclidean space $\mathbb{R}^d$, the tangent space at any point $x \in \mathbb{R}^d$ is $\mathbb{R}^d$ itself. A characterization of the gradient vector $\nabla f(x)$ is a (tangent) vector in $T_x\mathbb{R}^d = \mathbb{R}^d$ such that, for any other vector $v \in T_x\mathbb{R}^d = \mathbb{R}^d$, the directional derivative w.r.t. v is given as the inner product of v and $\nabla f(x)$:

$$\nabla_v f(x) = \langle \nabla f(x), v \rangle.$$

This characterization is generalized as follows. If $(c_t)_{t\in[0,1]}$ is a curve in $\mathbb{R}^d$, then the derivative of f along the curve c is given by

$$\partial_t f(c_t) = \langle \nabla f(c_t), \dot{c}_t \rangle$$

where $\dot{c}_t = \partial_t c_t$ is the speed vector of c at time t.

Definition 4.1. *Let $F : W_{2,ac}(\mathbb{R}^d) \to \mathbb{R}^d$ be a function. The Wasserstein gradient of F at $\mu \in W_{2,ac}(\mathbb{R}^d)$ is a vector $\mathbb{\nabla} F(\mu)$ in $T_\mu W_{2,ac}(\mathbb{R}^d)$ such that*

$$\partial_t|_{t=0} F(\mu_t) = \langle \mathbb{\nabla} F(\mu), v_0 \rangle_{L^2(\mu)}$$

for any curve μ_t in $W_{2,ac}(\mathbb{R}^d)$ and its tangent vector field v_t such that $\mu_0 = \mu$.

Wasserstein gradient of a function F can conveniently calculated by *first variation* of F.

Definition 4.2. *Let $F : W_{2,ac}(\mathbb{R}^d) \to \mathbb{R}$ and $\mu \in W_{2,ac}(\mathbb{R}^d)$. The first variation of F at μ is a function $\delta F(\mu) : \mathbb{R}^d \to \mathbb{R}$ such that*

$$\lim_{\epsilon \to 0} \frac{F(\mu + \epsilon\chi) - F(\mu)}{\epsilon} = \int \delta F(\mu) d\chi$$

for all measure χ such that $\mu + \epsilon\chi \in W_{2,ac}(\mathbb{R}^d)$ for all sufficiently small ϵ.

It can be shown that if the first variation exists, we have a generalized identity. For a curve μ_t in $W_{2,ac}(\mathbb{R}^d)$, using the idea of $\mu_{t+\epsilon} \approx \mu_t + \epsilon\partial_t\mu_t$, it holds that

$$\partial_t F(\mu_t) = \int \delta F(\mu_t)\partial_t\mu_t. \tag{8}$$

This equation leads to the following proposition.

Proposition 4.3. *Let $F : W_{2,ac}(\mathbb{R}^d) \to \mathbb{R}$ and $\mu \in W_{2,ac}(\mathbb{R}^d)$. Then,*

$$\mathbb{\nabla} F(\mu) = \nabla\delta F(\mu).$$

Proof. Let (μ_t, v_t) be any pair that solves the continuity equation.

$$\begin{aligned}
\langle \mathbb{\nabla} F(\mu), v_0\rangle_{L^2(\mu)} &= \partial_t|_{t=0} F(\mu_t) \\
&= \int \delta F(\mu_t)\partial_t\mu_t|_{t=0} \\
&= -\int \delta F(\mu_t)\mathrm{div}(\mu_t v_t)dx|_{t=0} \\
&= \int \langle \nabla\delta F(\mu_t), \mu_t v_t\rangle dx|_{t=0} \\
&= \langle \nabla\delta F(\mu), v_0\rangle_{L^2(\mu)}.
\end{aligned}$$

□

The following examples will be useful for computing Wasserstein gradient of KL divergence $KL(\cdot\|\pi)$ for some fixed π in a later section.

Example 4.4. *Let the functional is given in the form $F(\mu) = \int V d\mu$ for some function $V : \mathbb{R}^d \to \mathbb{R}$. Then $\partial_t F(\mu_t) = \int V\partial_t\mu_t$, and by (8), $\delta F(\mu_t) = V$ (constant over time). Therefore, $\mathbb{\nabla} F(\mu) = \nabla\delta F(\mu) = \nabla V$.*

Example 4.5. *Let $F(\mu) = \int U(\mu(x))dx$ for some function $U : \mathbb{R}_+ \to \mathbb{R}$. For example, the entropy functional $U(x) = x\log x$. Then, $\partial_t F(\mu_t) = \int U'(\mu_t)\partial_t\mu_t$, and by (8), $\delta F(\mu) = U' \circ \mu$. Therefore, $\mathbb{\nabla} F(\mu) = \nabla\delta F(\mu) = \nabla\log\mu$.*

Finally, a Wasserstein gradient flow of a functional F is a curve μ_t in $W_{2,ac}(\mathbb{R}^d)$ whose tangent vector is the negative of the Wasserstein gradient of F at μ_t at each time t. Precisely,

Definition 4.6 (Wasserstein gradient flow)**.** *A Wasserstein gradient flow of a functional $F : W_{2,ac}(\mathbb{R}^d) \to \mathbb{R}$ is a curve $(\mu_t)_{t\geq 0}$ in $W_{2,ac}(\mathbb{R}^d)$ that solves the PDE*

$$\partial_t\mu_t = div(\mu_t \mathbb{\nabla} F(\mu_t)).$$

5 Bures-Wasserstein Gradient

Definition 5.1. *The Bures–Wasserstein space* $\mathrm{BW}(\mathbb{R}^d)$ *is the space of non-degenerate Gaussians on $\mathbb{R}^d$, equipped with the Wasserstein metric.*

The purpose of this section is to characterize the Wasserstein gradient flow when restricted to this Gaussian submanifold. This formulation plays a central role in Gaussian variational inference (GVI), which we will introduce in Chapter 7. To develop a theory of gradient flows in Bures-Wasserstein space $\mathrm{BW}(\mathbb{R}^d)$, we need conditions for vector fields v_t that ensures the evolution of a non-degenerate Gaussian distribution μ along v_t remains a non-degenerate Gaussian distribution.

In the class, we showed that the (unique) optimal transport map between two non-degenerate Gaussian distributions is a non-degenerate affine map. More precisely, the optimal transport map can be expressed as follows (Chewi et al., 2024, Example 1.19).

Fact 5.2. *The unique optimal transport map from $N(m_1, \Sigma_1)$ to $N(m_2, \Sigma_2)$, where Σ_1, Σ_2 are non-singular, is given as follows.*

$$T(x) = \Sigma_1^{-1/2}\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\Sigma_1^{-1/2}(x - m_1) + m_2.$$

According to Theorem 3.1, the tangent space of $\mathrm{BW}(\mathbb{R}^d)$ is defined as

$$T_\mu \mathrm{BW}(\mathbb{R}^d) = \left\{x \mapsto Ax + a : A \in \mathbf{S}^d, a \in \mathbb{R}^d\right\}.$$

Here A is assumed to be symmetric to avoid having multiple tangent vectors that define equivalent transports.

Now the Bures-Wasserstein gradient $\nabla_{BW}F$ is defined similarly as in Theorem 3.1, but restricted to the Bures-Wasserstein tangent space.

Definition 5.3. *The Bures-Wasserstein gradient at μ is defined as a vector $\nabla_{BW}F(\mu)$ in $T_\mu \mathrm{BW}(\mathbb{R}^d)$ such that*

$$\partial_t|_{t=0}F(\mu_t) = \langle \nabla_{BW}F(\mu), v_0\rangle_{L^2(\mu)}$$

for any curve μ_t in $\mathrm{BW}(\mathbb{R}^d)$ such that $\mu_0 = \mu$ and its tangent vector field v_t.

Although we can compute BW gradient from scratch as in the previous section, a more convenient approach is to leverage the equation $\mathbb{\nabla}F(\mu) = \nabla\delta F(\mu)$. It holds that

$$\langle \nabla_{BW}F(\mu), v_0\rangle_{L^2(\mu)} = \partial_t|_{t=0}F(\mu_t) = \langle \mathbb{\nabla}F(\mu), v_0\rangle_{L^2(\mu)} = \langle \nabla\delta F(\mu), v_0\rangle_{L^2(\mu)} \tag{9}$$

for all $v_0 \in T_\mu \mathrm{BW}(\mathbb{R}^d)$. The affine map $\nabla_{BW}F(\mu)$ that satisfies (9) for all $v_0 \in T_\mu \mathrm{BW}(\mathbb{R}^d)$ is given by the following form.

Proposition 5.4. *Let $F : W_{2,ac}(\mathbb{R}^d) \to \mathbb{R}$ be a functional and $\mu \sim \mathcal{N}(m, \Sigma) \in \mathrm{BW}(\mathbb{R}^d)$. The Bures-Wasserstein gradient of F at $\mu \in \mathrm{BW}(\mathbb{R}^d)$ is the map*

$$x \mapsto \left(\int \nabla^2\delta F(\mu)d\mu\right)(x - m) + \int \nabla\delta F(\mu)d\mu.$$

Sketch of proof. Recall that

$$\langle v, w\rangle_{L^2(\mu)} = \int \langle v(x), w(x)\rangle \mu(x)dx = \mathbb{E}_{X\sim\mu}\langle v(X), w(X)\rangle.$$

With the above notations, one finds parameters $A \in \mathbf{S}^d, a \in \mathbb{R}^d$ for

$$\nabla_{BW}F(\mu) = A(\cdot - m) + a$$

such that

$$\mathbb{E}\langle A(X-m)+a, \tilde{A}(X-m)+\tilde{a}\rangle = \mathbb{E}\langle \nabla\delta F(\mu)(X), \tilde{A}(X-m)+\tilde{a}\rangle \tag{10}$$

for any $\tilde{A} \in \mathbf{S}^d, \tilde{a} \in \mathbb{R}^d$.

$$\begin{aligned}
\mathbb{E}\langle A(X-m), \tilde{A}(X-m)\rangle &= \mathbb{E}\left(A(X-m)\right)^\top \tilde{A}(X-m) \\
&= \mathbb{E}\left(\operatorname{tr}\left(\tilde{A}(X-m)(X-m)^\top A^\top\right)\right) \\
&= \operatorname{tr}\left(\tilde{A}\Sigma A^\top\right) \\
&= \langle A, \Sigma\tilde{A}\rangle.
\end{aligned}$$

In the last equality, we use brackets for inner product of matrices,

$$\langle A, B\rangle = \sum_{i,j} A_{ij}B_{ij} = \operatorname{tr}(A^\top B) = \operatorname{tr}(AB^\top).$$

Hence

$$(\text{LHS of (10)}) = \langle A, \Sigma\tilde{A}\rangle + \langle a, \tilde{a}\rangle. \tag{11}$$

Some calculus (see Chewi et al., 2024, Theorem 5.21 for detailed calculation) shows that

$$(\text{RHS of (10)}) = \left\langle \int \nabla^2\delta F(\mu)d\mu, \Sigma\tilde{A}\right\rangle + \left\langle \int \nabla\delta F(\mu)d\mu, \tilde{a}\right\rangle. \tag{12}$$

Setting (11) equals to (12) gives the proposition. □

Finally, we briefly discuss the convexity of the BW space. Since Gaussian measures are closed under Wasserstein geodesics with the optimal transport map between any $\mu_0, \mu_1 \in \mathrm{BW}(\mathbb{R}^d)$ being affine, the geodesic between any two elements in $\mathrm{BW}(\mathbb{R}^d)$ remains in $\mathrm{BW}(\mathbb{R}^d)$. This immediately implies the following convexity result.

Proposition 5.5. $\mathrm{BW}(\mathbb{R}^d) \subset W_{2,ac}(\mathbb{R}^d)$ *is geodesically convex.*

6 Convergence Rate of Gradient Flows and Geodesic Convexity of KL divergence

This section studies (i) the convergence rate of gradient flows and (ii) the geodesic convexity of the KL divergence over the Wasserstein space.

As discussed in Section 4, the Wasserstein gradient flow of a functional F over $W_{2,ac}(\mathbb{R}^d)$ is a curve of measures $(\mu_t)_{t\geq 0}$ whose tangent vector at time t is given by $v_t = -\mathbb{\nabla} F(\mu_t)$. This vector field governs the evolution of $(\mu_t)_{t\geq 0}$ through the continuity equation. Based on Definition 4.1, the time derivative of the functional along the flow satisfies

$$\partial_t F(\mu_t) = \langle \mathbb{\nabla} F(\mu_t), v_t\rangle_{\mu_t} = -\|\mathbb{\nabla} F(\mu_t)\|_{\mu_t}^2. \tag{13}$$

Before we proceed, let us equip ourselves with some definitions and properties from Riemannian geometry.

Definition 6.1. $C \subset \mathcal{M}$ *is geodesically convex if for all* $p_0, p_1 \in C$*, all geodesics joining* p_0 *to* p_1 *also lie in* C*.*

Definition 6.2. *Given $\alpha \in \mathbb{R}$ and a manifold $\mathcal{M}$, a function $f : \mathcal{M} \to \mathbb{R}$ is called α-geodesically convex if*

$$f(p_t) \leq (1-t)f(p_0) + tf(p_1) - \frac{\alpha t(1-t)}{2} d^2(p_0, p_1), \quad \forall t \in [0,1],$$

for all geodesics $(p_t)_{t\in[0,1]}$.

Proposition 6.3. *The following are equivalent for a function $f : \mathcal{M} \to \mathbb{R}$:*

1. *f is α-geodesically convex.*
2. *(First order condition) For all $p, q \in \mathcal{M}$,*
$$f(q) \geq f(p)\langle \nabla f(p), \log_p(q)\rangle_p + \frac{\alpha}{2} d^2(p,q), \tag{14}$$
where ∇f, the Riemannian gradient, defined so that for all curves $(p_t)_{t\geq 0}$, $\nabla f(p_t) \in T_{p_t}\mathcal{M}$, satisfies $\partial_t f(p_t) = \langle \nabla f(p_t), \dot{p}_t\rangle_{p_t}$.
3. *(Second-order condition) For all $p \in \mathcal{M}$, $v \in T_p\mathcal{M}$,*
$$\nabla^2 f(p)[v,v] \geq \alpha \|v\|_p^2, \tag{15}$$
where $\nabla^2 f$ is the Riemannian Hessian, defined as $\nabla^2 f(p)[v,v] := \partial_t^2 f(p_t)\big|_{t=0}$, with $(p_t)_{t\in[0,1]}$ being the geodesic satisfying $p_0 = p$ and $\dot{p_0} = v$.

The derivation of the convergence rate for gradient flows primarily relies on Grönwall's inequality stated below:

Lemma 6.4 (Grönwall's inequality)**.** *Let $c \in \mathbb{R}$. Let $\phi : [0,T] \to \mathbb{R}$ be differentiable, satisfying $\dot{\phi}(t) \leq c\phi(t)$ for all $t \in [0,T]$. Then,*

$$\phi(t) \leq \phi(0)\exp(ct) \quad \forall t \in [0,T].$$

Proof. Let $g(t) = \exp(-ct)\phi(t)$. By the assumption,

$$\partial_t g(t) = \exp(-ct)[-c\phi(t) + \dot{\phi}(t)] \leq 0.$$

Thus, $g(t) = \exp(-ct)\phi(t) \leq g(0) = \phi(0)$. □

Note that if the inequality in Lemma 6.4 were an equality, the solution would be $\phi(t) = \phi(0)\exp(ct)$. In general, the lemma tells us that the solution to a differential inequality is bounded by the solution to the corresponding differential equation. In our context, a differential inequality of this form can be introduced as the PL inequality condition.

Definition 6.5 (Polyak-Lojasiewicz (PL) inequality)**.** *We say that $F : WS \to \mathbb{R}$ satisfies a PL inequality with constant alpha > 0 if for all $\mu \in W_{2,ac}(\mathbb{R}^d)$,*

$$\|\nabla\!\!\!\nabla F(\mu)\|_\mu^2 \geq 2\alpha(F(\mu) - \inf F).$$

Combining the PL inequality with the identity in (13) yields

$$\partial_t(F(\mu_t) - \inf F) \leq -2\alpha(F(\mu_t) - \inf F),$$

which can be rewritten as a differential inequality $\dot{\phi}(t) \leq -2\alpha\phi(t)$, with $\phi(t) := F(\mu_t) - \inf F$. By Lemma 6.4, this implies exponential convergence of the functional values along the gradient flow. Finally, since F being α-geodesically convex implies that it satisfies the PL inequality with α, Corollary 6.6 obtain the exponential convergence rate.

Corollary 6.6. *Let $F : W_{2,ac}(\mathbb{R}^d) \to \mathbb{R}$ be α-geodesically convex. Then, along the Wasserstein gradient flow $(\mu_t)_{t\geq 0}$ for F, it holds*

$$F(\mu_t) - \inf F \leq e^{-2\alpha t}(F(\mu_0) - \inf F).$$

Next, we introduce the Variational Inference (VI) problem and discuss the geodesic convexity of its objective. Before focusing on the Gaussian VI in Section 7, we proceed with the general VI formulation: the goal is to approximate a certain probability measure π with an element of a simpler class Q of probability measures by solving the optimization problem:

$$q^* = \arg\min_{q \in Q} \mathrm{KL}(q \| \pi). \tag{16}$$

Our goal here is to show the geodesic convexity of the KL divergence. We assume that π admits a density of the form $\pi \propto \exp(-V)$, where $V : \mathbb{R}^d \to \mathbb{R}$ is referred to as the potential function. The corresponding objective functional becomes:

$$F(\mu) := \mathrm{KL}(\mu \| \pi) = \int \mu \log \frac{\mu}{\pi} = \underbrace{\int V \, d\mu}_{=:\mathcal{V}(\mu)} + \underbrace{\int \mu \log \mu}_{=:\mathcal{H}(\mu)} + \text{const}. \tag{17}$$

Recall to the Example 4.4 and 4.5, we have $\mathbb{\nabla}\mathcal{V}(\mu) = \nabla V$ and $\mathbb{\nabla}\mathcal{H}(\mu) = \nabla \log \mu$, and thus $\mathbb{\nabla} F(\mu) = \nabla \log \mu + \nabla V = \nabla \log(\mu/\pi)$.

We can consider the convexity of two functionals $\mathcal{V}$ and $\mathcal{H}$ introduced in (17) separately. Specifically, Theorem 6.7 states the convexity of $\mathcal{V}$, while Theorem 6.8 states the convexity of $\mathcal{H}$.

Theorem 6.7. *Suppose $V : \mathbb{R}^d \to \mathbb{R}$ is α-convex on $\mathbb{R}^d$. Then, $\mathcal{V}(\mu) := \int V d\mu$ is α-geodesically convex on $W_{2,ac}(\mathbb{R}^d)$.*

Proof. We will use the second-order condition of (15). Let $X_t \sim \mu_t$ for $t \in [0,1]$, where $(\mu_t)_{t\in[0,1]}$ is a Wasserstein geodesic. Then, $X_t = (1-t)X_0 + tT(X_0)$ and $\dot{X}_t = T(X_0) - X_0$, where T is the optimal transport map from μ_0 to μ_1. We compute

$$\begin{aligned}
\mathbb{\nabla}^2\mathcal{V}(\mu_0)[T - \mathrm{id}, T - \mathrm{id}] &= \partial_t^2 \mathcal{V}(\mu_t)\big|_{t=0} = \partial_t^2 \mathbb{E} V(X_t)\big|_{t=0} \\
&= \mathbb{E}(\dot{X}_t^{\top} \nabla^2 V(X_t) \dot{X}_t) \\
&= \mathbb{E}\langle T(X_0) - X_0, \nabla^2 V(X_0)\,(T(X_0) - X_0)\rangle \\
&\geq \alpha \, \mathbb{E}\, \|T(X_0) - X_0\|^2 \\
&= \alpha \, \|T - \mathrm{id}\|_{\mu_0}^2 ,
\end{aligned}$$

where we used the assumption $\nabla^2 V \succeq \alpha I$. □

Theorem 6.8. *The entropic functional $\mathcal{H}(\mu) := \int \mu \log \mu$ is geodesically convex on $W_{2,ac}(\mathbb{R}^d)$.*

Proof. Let $(\mu_t)_{t\in[0,1]}$ be a Wasserstein geodesic and let $T_t := (1-t)\,\mathrm{id} + t\,T$, so that $\mu_t = (T_t)_{\#}\mu_0$. Then,

$$\begin{aligned}\mathcal{H}(\mu_t) &= \int (\log \mu_t)\,\mathrm{d}\mu_t = \int \log(\mu_t \circ T_t)\,\mathrm{d}\mu_0 = \int \log\left(\frac{\mu_0}{\det \nabla T_t}\right)\mathrm{d}\mu_0 \\ &= \mathcal{H}(\mu_0) - \int \log\det \nabla T_t\,\mathrm{d}\mu_0,\end{aligned}$$

where the third equality holds from the change of variable.

Note that $-\log\det$ is convex over the positive definite cone, and $t \mapsto \nabla T_t$ is affine. Therefore, the composition $t \mapsto -\log\det \nabla T_t$ is convex. Thus, $t \mapsto \mathcal{H}(\mu_t)$ is convex. □

Having established that both $\mathcal{V}(\mu)$ and $\mathcal{H}(\mu)$ are geodesically convex, we now conclude that the KL divergence functional $F(\mu) := \mathrm{KL}(\mu\|\pi)$ is geodesically convex on $W_{2,ac}(\mathbb{R}^d)$. This is summarized in the following corollary, which directly follows from Theorems 6.7 and 6.8.

Corollary 6.9. *Let $\pi \propto \exp(-V)$ be a density, where $V : \mathbb{R}^d \to \mathbb{R}$ is α-convex. Then, the KL divergence functional $\mathcal{F} := \mathrm{KL}(\cdot\|\pi)$ is α-geodesically convex on $W_{2,ac}(\mathbb{R}^d)$.*

In practical variational inference settings, we are often interested in minimizing the KL divergence over a restricted class $Q \subseteq W_{2,ac}(\mathbb{R}^d)$. If this class Q is itself geodesically convex, the geodesic convexity of F naturally extends to this constrained setting as well.

Corollary 6.10. *Let $\pi \propto \exp(-V)$ be a density with α-convex potential V, and let $Q \subseteq W_{2,ac}(\mathbb{R}^d)$ be geodesically convex. Then, the KL divergence is α-geodesically convex over Q, and the solution q^* to the VI problem in* (16) *is unique.*

Furthermore, similarly to what we have used in Corollary 6.6, the fact that α-geodesic convexity implies that the PL inequality allows us to deduce an exponential convergence rate for the Wasserstein gradient flow of the KL divergence.

Corollary 6.11. *Let $\pi \propto \exp(-V)$ be a density on $\mathbb{R}^d$, where V is α-convex, and let $Q \subseteq W_{2,ac}(\mathbb{R}^d)$ be geodesically convex. Then, the Wasserstein gradient flow $(q_t)_{t\geq 0}$ of $\mathrm{KL}(\cdot\|\pi)$, constrained to lie in Q, satisfies*

$$\mathrm{KL}(q_t\|\pi) - \mathrm{KL}(q^*\|\pi) \leq e^{-2\alpha t}\left(\mathrm{KL}(q_0\|\pi) - \mathrm{KL}(q^*\|\pi)\right),$$

where q^* is the minimizer of $\mathrm{KL}(\cdot\|\pi)$ in Q.

7 Gaussian Variation Inference via Otto Calculus

In this section, we consider the setting where the variational family Q consists of Gaussian distributions. This leads to the formulation known as Gaussian Variational Inference (GVI), which we have already introduced in (2). To solve the GVI problem, we follow the Wasserstein gradient flow constrained to the Bures-Wassersten (BW) space of Gaussian measures. Since the BW space is geodesically convex (Proposition 5.5), Corollary 6.11 applies.

We aim here to compute the explicit form of the BW gradient of the KL divergence functional. For $\mathcal{F} = \mathrm{KL}(\cdot\|\pi)$, where $\pi \propto \exp(-V)$, we have

$$\nabla\delta\mathcal{F}(q) = \nabla V + \nabla \log q, \quad \text{and} \quad \nabla^2\delta\mathcal{F}(q) = \nabla^2 V + \nabla^2 \log q. \tag{18}$$

Using Theorem 5.4, the BW gradient can be written as:

$$\begin{aligned}\nabla_{\mathrm{BW}}\mathcal{F}(q)(x) &= \left(\int (\nabla^2 V + \nabla^2 \log q)\, dq\right)(x - m_q) + \int \nabla V\, dq + \int \nabla \log q\, dq \\ &= \left(\int \nabla^2 V\, dq - \Sigma_q^{-1}\right)(x - m_q) + \int \nabla V\, dq,\end{aligned}$$

where m_q and Σ_q denote the mean and covariance of q, respectively.

Setting the BW gradient to zero solves the GVI problem and ensures uniqueness under convexity.

Proposition 7.1. *Suppose that $\pi \propto \exp(-V)$, where V is α-convex for some $\alpha > 0$. Then, the unique minimizer $q^\star$ in* (2) *is characterized by the conditions*

$$\int \nabla V\, \mathrm{d}q_\star = 0 \quad \textit{and} \quad \int \nabla^2 V\, \mathrm{d}q_\star = \Sigma_{q_\star}^{-1}.$$

We can now write down the BW gradient flow.

Theorem 7.2. *The BW gradient flow of the functional F is the curve $(\mu_t = N(m_t, \Sigma_t))_{t\geq 0}$, where*

$$\dot{m}_t = -\mathbb{E}\nabla\delta F(\mu_t)(X_t), \tag{19}$$

$$\dot{\Sigma}_t = -\mathbb{E}\nabla^2\delta F(\mu_t)(X_t)\,\Sigma_t - \Sigma_t\,\mathbb{E}\nabla^2\delta F(\mu_t)(X_t), \tag{20}$$

and $X_t \sim \mu_t$.

Proof. Based on the Proposition 5.4, the BW gradient flow can be represented as

$$\begin{aligned}\dot{X}_t &= -\nabla_{BW} F(\mu_t)(X_t) \\ &= -\left(\int \nabla^2\delta F(\mu_t) d\mu_t\right)(X_t - m_t) - \int \nabla\delta F(\mu_t) d\mu_t, \\ \dot{m}_t &= \mathbb{E}\dot{X}_t = -\int \nabla\delta F(\mu_t) d\mu_t, \\ \dot{\Sigma}_t &= \partial_t \mathbb{E}((X_t - m_t)(X_t - m_t)^\top) \\ &= \mathbb{E}(\dot{X}_t (X_t - m_t)^\top) + \mathbb{E}((X_t - m_t)\dot{X}_t^{\top}) \\ &= -\left(\int \nabla^2\delta F(\mu_t) d\mu_t\right)\Sigma_t - \Sigma_t\left(\int \nabla^2\delta F(\mu_t) d\mu_t\right).\end{aligned}$$

□

We now apply Theorem 7.2 to the GVI problem, by plugging in (18). Note that there is a slight simplification since the covariance matrix Σ_t cancels with the Hessian of the first variation of the entropy.

Theorem 7.3. *The BW gradient flow of the functional* $\mathrm{KL}(\cdot\|\pi)$, *where $\pi \propto \exp(-V)$, is the curve $(q_t = \mathcal{N}(m_t, \Sigma_t))_{t\geq 0}$, where*

$$\begin{aligned}\dot{m}_t &= -\mathbb{E}\nabla V(X_t), \\ \dot{\Sigma}_t &= -\mathbb{E}\nabla^2 V(X_t)\Sigma_t - \Sigma_t\mathbb{E}\nabla^2 V(X_t) + 2I,\end{aligned}$$

and $X_t \sim q_t$.

References

Ambrosio, L., Gigli, N., and Savaré, G. (2008). *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media.

Chewi, S., Niles-Weed, J., and Rigollet, P. (2024). Statistical optimal transport. *arXiv preprint arXiv:2407.18163*.